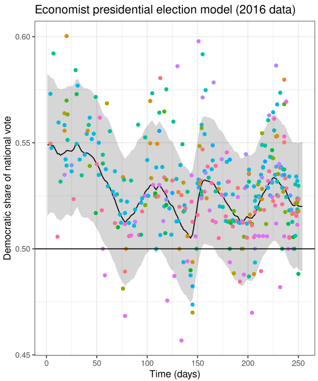
Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

April 2024

Theory and Foundations of Statistics in the Era of Big Data

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

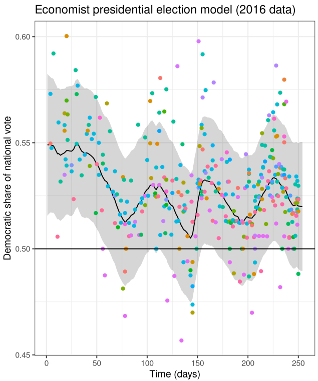
Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

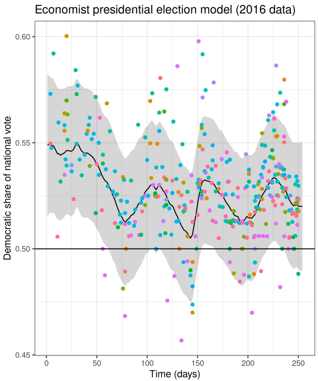
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

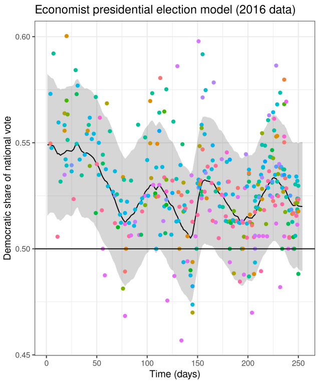
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time
- Is there high variability under re-sampling? [Huggins and Miller, 2023]
  Re-fit with bootstrap samples of data

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

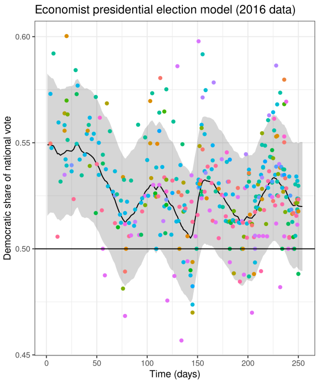
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time
- Is there high variability under re-sampling? [Huggins and Miller, 2023]
  Re-fit with bootstrap samples of data
- Are a small proportion (1%) of polls highly influential? [Broderick et al., 2020]
  Re-fit with sets of all 1% of datapoints removed

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)]$.

Some typical model checking tasks:

- How well are polls fit under cross-validation (CV)? [Vehtari and Ojanen, 2012]
  Re-fit with data points removed one at a time
- Is there high variability under re-sampling? [Huggins and Miller, 2023]
  Re-fit with bootstrap samples of data
- Are a small proportion (1%) of polls highly influential? [Broderick et al., 2020]
  Re-fit with sets of all 1% of datapoints removed

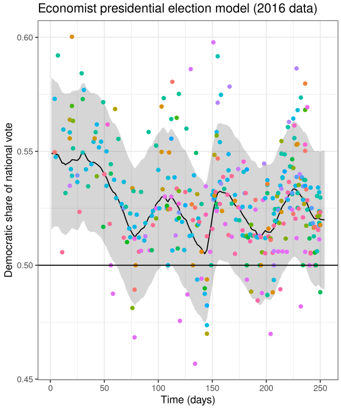
Problem: Each MCMC run takes about 10 hours (Stan, six cores).

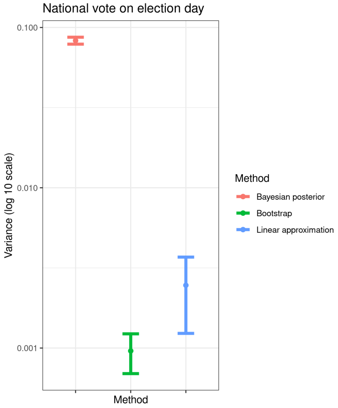
Results

We propose: Use posterior draws based on the full data, to form a linear approximation to *data reweightings*.

Results

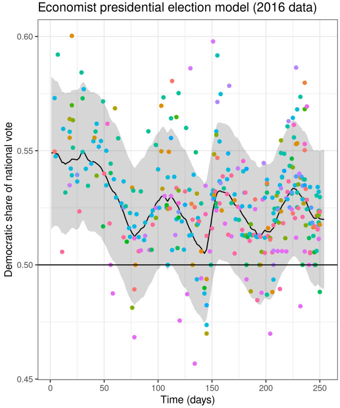
We propose: Use posterior draws based on the full data, to form a linear approximation to *data reweightings*.

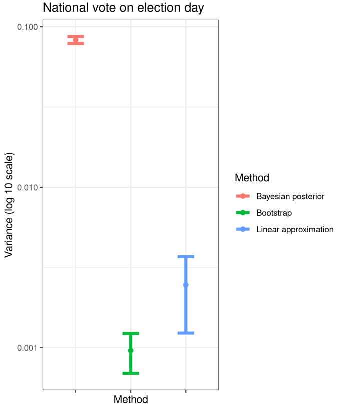




Results

We propose: Use posterior draws based on the full data, to form a linear approximation to *data reweightings*.





Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate, even as $N \to \infty$

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate, even as $N \to \infty$
- A trick question, and some implications of different weightings.

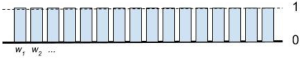
Data re-weighting.

Data re-weighting.

Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:

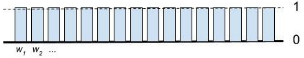


Data re-weighting.

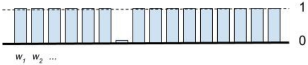
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:



Leave-one-out weights:

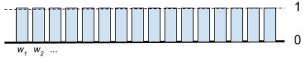


Data re-weighting.

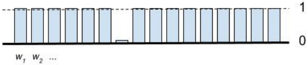
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

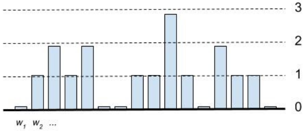
Original weights:



Leave-one-out weights:



Bootstrap weights:

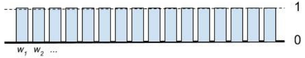


Data re-weighting.

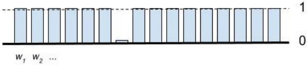
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

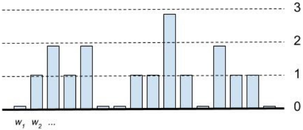
Original weights:

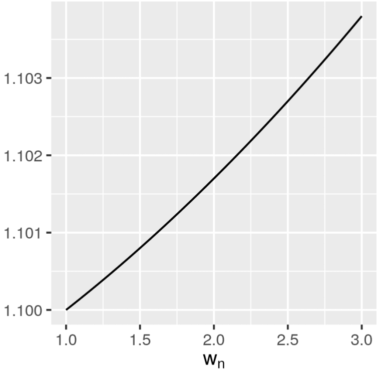


Leave-one-out weights:



Bootstrap weights:



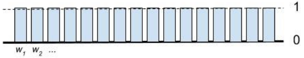


Data re-weighting.

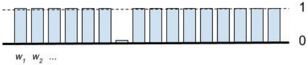
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

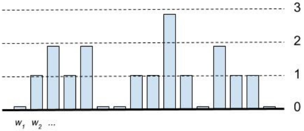
Original weights:

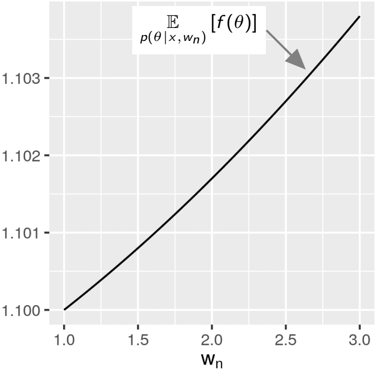


Leave-one-out weights:



Bootstrap weights:



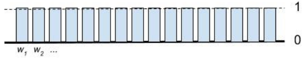


Data re-weighting.

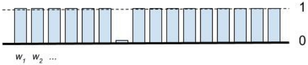
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

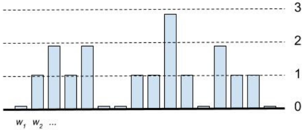
Original weights:

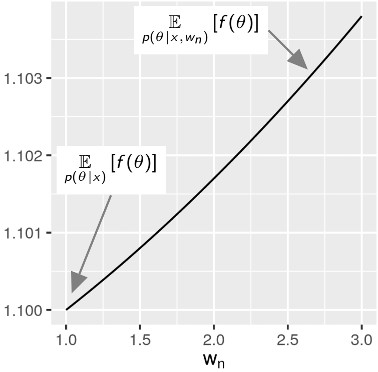


Leave-one-out weights:



Bootstrap weights:



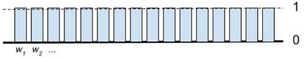


Data re-weighting.

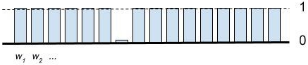
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

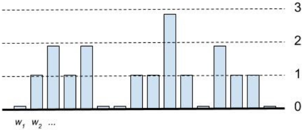
Original weights:

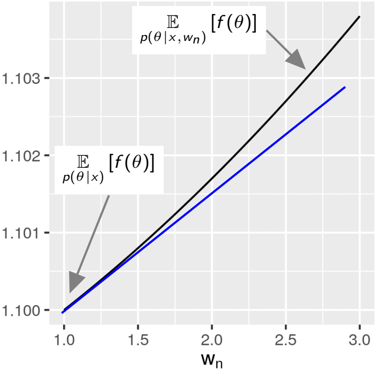


Leave-one-out weights:



Bootstrap weights:



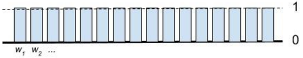


Data re-weighting.

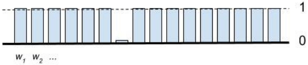
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

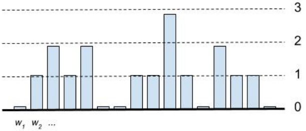
Original weights:

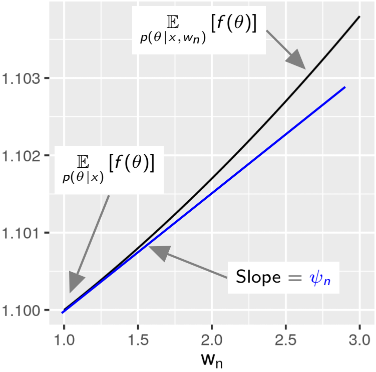


Leave-one-out weights:



Bootstrap weights:



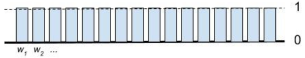


Data re-weighting.

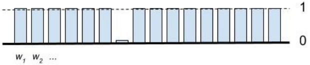
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

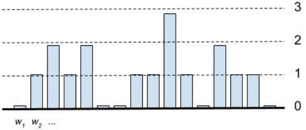
Original weights:

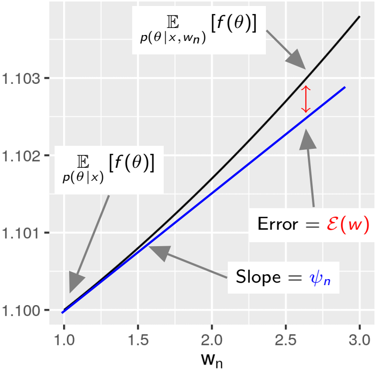


Leave-one-out weights:



Bootstrap weights:



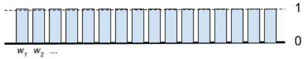


Data re-weighting.

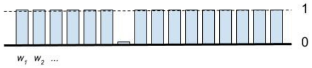
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

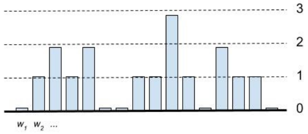
Original weights:

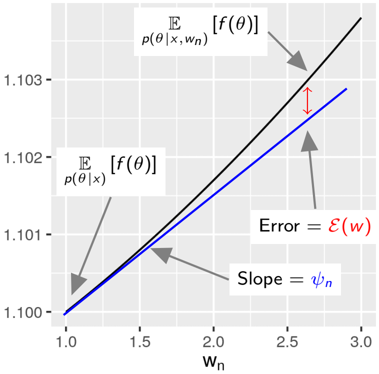


Leave-one-out weights:



Bootstrap weights:





The re-scaled slope $N\psi_n$ is known as the "influence function" at data point x_n.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})}[f(\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] - \psi_n$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})}[f(\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] - \psi_n$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \mathrm{Multinomial}(N, N^{-1})$.

$$\text{Bootstrap variance} = \mathop{\mathrm{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|x,w)}[f(\theta)]\right) \approx \frac{1}{N^2}\sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})}[f(\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] - \psi_n$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \mathrm{Multinomial}(N, N^{-1})$.

$$\text{Bootstrap variance} = \mathop{\mathrm{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|x,w)}[f(\theta)]\right) \approx \frac{1}{N^2}\sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2$$

Influential subsets: Approximate maximum influence perturbation (AMIP).

Let $W_{(-K)}$ denote weights leaving out K points.

$$\max_{w \in W_{(-K)}} \left(\mathop{\mathbb{E}}_{p(\theta|x,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)]\right) \approx -\sum_{n=1}^{K} \psi_{(n)}.$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar mean posterior–mean zero (e.g., $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$).

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}(w_n) = \frac{1}{2}\underbrace{\underset{p(\theta|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

$= O_p(N^{-1})$ under a BCLT

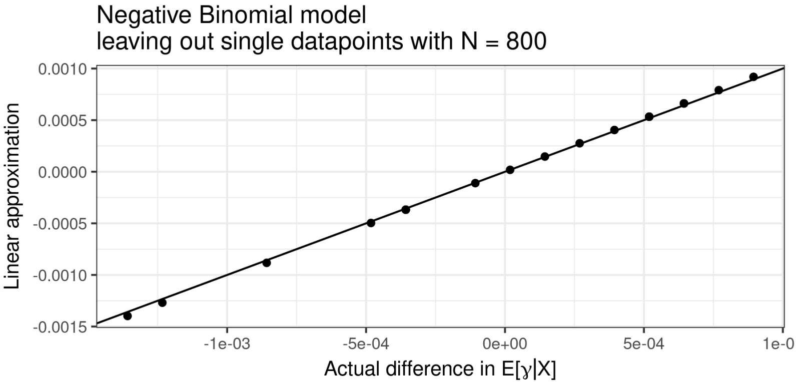
$= O_p(N^{-2})$ under a BCLT

$\Rightarrow$ The map $w_n \mapsto N\left(\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]\right)$ becomes linear as $N \to \infty$.

(See [Kass et al., 1990] for a *particular weight*, [?] for a kind of uniform convergence over datapoints.)

Low dimensional problems

Example: **Negative binomial models with an unkown parameter γ.**



The map $w_n \mapsto N \left(\underset{p(\gamma|X, w_n)}{\mathbb{E}} [\gamma] - \underset{p(\gamma|X)}{\mathbb{E}} [\gamma] \right)$ becomes linear as $N \to \infty$.

High dimensional problems

What about when the posterior doesn't obey a BCLT?

Suppose that $p(\lambda|X)$ does not concentrate.

(E.g., N is small, λ grows in dimension or X is uninformative.)

High dimensional problems

What about when the posterior doesn't obey a BCLT?

Suppose that $p(\lambda|X)$ does not concentrate.

(E.g., N is small, λ grows in dimension or X is uninformative.)

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{p(\lambda|X,w_n)}[f(\lambda)] - \mathop{\mathbb{E}}_{p(\lambda|X)}[f(\lambda)] \\
&= \psi_n(w_n - 1) \qquad\qquad\qquad\qquad + \mathcal{E}(w_n) \\
&= \underbrace{\mathop{\mathbb{E}}_{p(\lambda|X)}\left[\bar{f}(\lambda)\bar{\ell}_n(\lambda)\right](w_n - 1)}_{O_p(1)} \qquad + \frac{1}{2}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|X,\tilde{w}_n)}\left[\bar{f}(\lambda)\bar{\ell}_n(\lambda)\bar{\ell}_n(\lambda)\right](w_n - 1)^2}_{O_p(1)}.
\end{aligned}
$$

The error is of the same order as the slope.

The map $w_n \mapsto \mathop{\mathbb{E}}_{p(\lambda|X,w_n)}[f(\lambda)]$ is nonlinear in general.

High dimensional problems

What about when the posterior doesn't obey a BCLT?

Suppose that $p(\lambda|X)$ does not concentrate.

(E.g., N is small, λ grows in dimension or X is uninformative.)

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{p(\lambda|X,w_n)}[f(\lambda)] - \mathop{\mathbb{E}}_{p(\lambda|X)}[f(\lambda)] \\
&= \psi_n(w_n - 1) \qquad\qquad + \mathcal{E}(w_n) \\
&= \underbrace{\mathop{\mathbb{E}}_{p(\lambda|X)}\left[\bar{f}(\lambda)\bar{\ell}_n(\lambda)\right](w_n - 1)}_{O_p(1)} \qquad + \underbrace{\frac{1}{2}\mathop{\mathbb{E}}_{p(\lambda|X,\tilde{w}_n)}\left[\bar{f}(\lambda)\bar{\ell}_n(\lambda)\bar{\ell}_n(\lambda)\right](w_n - 1)^2}_{O_p(1)}.
\end{aligned}
$$

The error is of the same order as the slope.

The map $w_n \mapsto \mathop{\mathbb{E}}_{p(\lambda|X,w_n)}[f(\lambda)]$ is nonlinear in general.

Can we save the approximation when *some* parameters concentrate?

Example: **Poisson model with random effects (REs) λ and fixed effects γ.**

High dimensional problems

Example: **Poisson model with random effects (REs) λ and fixed effects γ.**

If the observations per random effect remains bounded as $N \to \infty$, then

- Parameter λ ("local") grows in dimension with N.
- Parameter γ ("global") is finite-dimensional.
- Marginally $p(\lambda|X)$ does not concentrate.
- Marginally, $p(\gamma|X)$ concentrates.

High dimensional problems

Example: **Poisson model with random effects (REs) λ and fixed effects γ.**

If the observations per random effect remains bounded as $N \to \infty$, then

- Parameter λ ("local") grows in dimension with N.
- Parameter γ ("global") is finite-dimensional.
- Marginally $p(\lambda|X)$ does not concentrate.
- Marginally, $p(\gamma|X)$ concentrates.

Can we save the approximation when *some* parameters concentrate?
$\Rightarrow$ Does the residual vanish asymptotically for $w_n \mapsto \underset{p(\gamma|X, w_n)}{\mathbb{E}}[\gamma]$?

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\underset{p(\gamma,\lambda|X,w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] =$$

$$\psi_n(w_n - 1) \qquad\qquad + \mathcal{E}(w_n)$$

$$= \underset{p(\gamma,\lambda|X)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n - 1) \qquad + \frac{1}{2}\underset{p(\gamma,\lambda|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n - 1)^2$$

$$\psi_n = O_p(N^{-1}) \qquad\qquad \mathcal{E}(w_n) = O_p(N^{-1})$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$
\begin{aligned}
&\underset{p(\gamma,\lambda|X,w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] = \\
&\quad \psi_n(w_n - 1) && + \mathcal{E}(w_n) \\
&= \underset{p(\gamma,\lambda|X)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n - 1) && + \frac{1}{2}\underset{p(\gamma,\lambda|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n - 1)^2 \\
&= \underset{p(\gamma|X)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|\gamma,X)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\right](w_n - 1) && + \frac{1}{2}\underset{p(\gamma|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|X,\gamma,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\right](w_n - 1)^2
\end{aligned}
$$

$$\psi_n = O_p(N^{-1}) \qquad \mathcal{E}(w_n) = O_p(N^{-1})$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\underset{p(\gamma,\lambda|X,w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] =$$

$$\psi_n(w_n - 1) \qquad + \mathcal{E}(w_n)$$

$$= \underset{p(\gamma,\lambda|X)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n-1) \qquad + \frac{1}{2}\underset{p(\gamma,\lambda|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n-1)^2$$

$$= \underset{p(\gamma|X)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|\gamma,X)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\right](w_n-1) \qquad + \frac{1}{2}\underset{p(\gamma|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|X,\gamma,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\right](w_n-1)^2$$

$$= \underbrace{\underset{p(\gamma|X)}{\mathbb{E}}[\bar{\gamma}F_1(\gamma)]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n-1) \qquad + \frac{1}{2}\underbrace{\underset{p(\gamma|X,\tilde{w}_n)}{\mathbb{E}}[\bar{\gamma}F_2(\gamma)]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n-1)^2$$

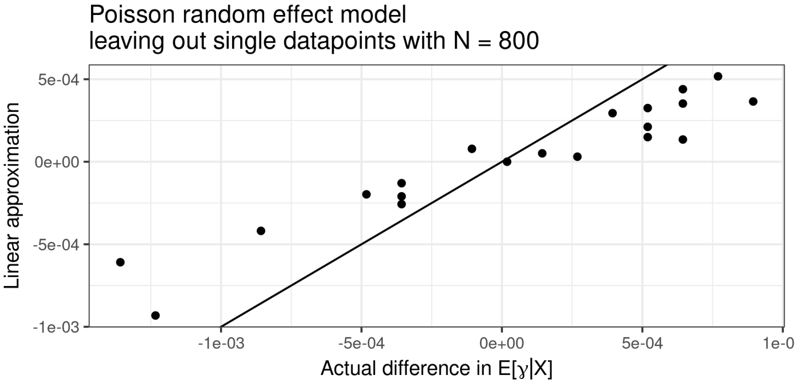
$$\Rightarrow \psi_n = O_p(N^{-1}) \qquad \mathcal{E}(w_n) = O_p(N^{-1})$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$
\begin{aligned}
&\underset{p(\gamma,\lambda|X,w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] = \\
&\quad \psi_n(w_n - 1) && + \mathcal{E}(w_n) \\
&= \underset{p(\gamma,\lambda|X)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n-1) && + \frac{1}{2}\underset{p(\gamma,\lambda|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n-1)^2 \\
&= \underset{p(\gamma|X)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|\gamma,X)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\right](w_n-1) && + \frac{1}{2}\underset{p(\gamma|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|X,\gamma,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\right](w_n-1)^2 \\
&= \underbrace{\underset{p(\gamma|X)}{\mathbb{E}}\left[\bar{\gamma}F_1(\gamma)\right]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n-1) && + \frac{1}{2}\underbrace{\underset{p(\gamma|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}F_2(\gamma)\right]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n-1)^2 \\
&\quad \Rightarrow \psi_n = O_p(N^{-1}) && \mathcal{E}(w_n) = O_p(N^{-1})
\end{aligned}
$$

The map $w_n \mapsto N\left(\underset{p(\gamma|X,w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma|X)}{\mathbb{E}}[\gamma]\right)$ remains non-linear as $N \to \infty$.


Poisson random effect model
leaving out single datapoints with N = 800
Linear approximation
5e-04
0e+00
-5e-04
-1e-03
-1e-03
-5e-04
0e+00
5e-04
1e-0
Actual difference in E[γ|X]

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X,w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

$$\log p(X|\gamma, w^m) = \sum_{n=1}^{N} w_n^m \log p(x_n|\gamma)$$

Poisson observations with random effects.

Asymptotically non-linear in w.

$$\log p(X|\gamma, \lambda, w^c) = \sum_{n=1}^{N} w_n^c \log p(x_n|\lambda, \gamma)$$

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X,w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

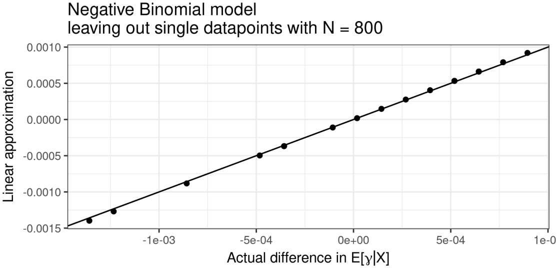
Trick question! We weight a log likelihood contribution, not a datapoint.

The two weightings are not equivalent in general.

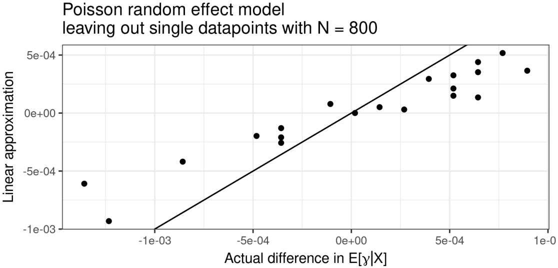
Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Approximation based on $\log p(x_n|\gamma)$.



Approximation based on $\log p(x_n|\gamma, \lambda)$.

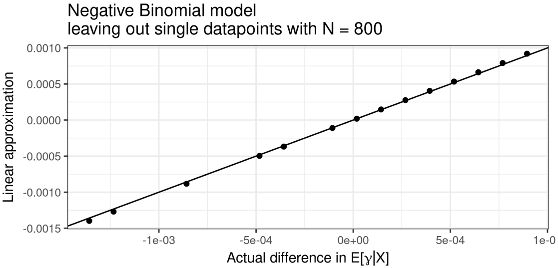


Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

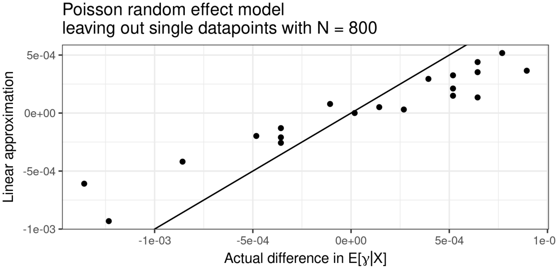
Approximation based on $\log p(x_n|\gamma)$.

Not computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$ in general.



Approximation based on $\log p(x_n|\gamma, \lambda)$.

Computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$.

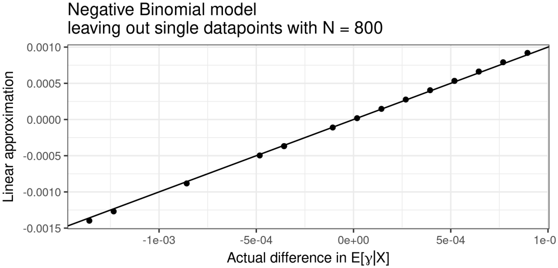


Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Approximation based on $\log p(x_n|\gamma)$.

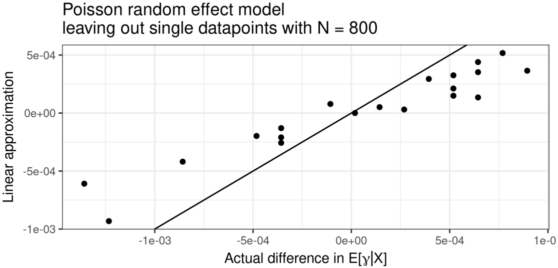
Not computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$ in general.



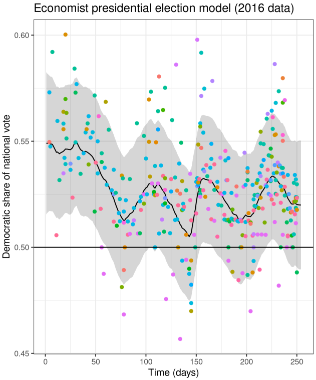
Approximation based on $\log p(x_n|\gamma, \lambda)$.

Computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$.

May still be useful when $p(\lambda|X)$ is *somewhat* concentrated.

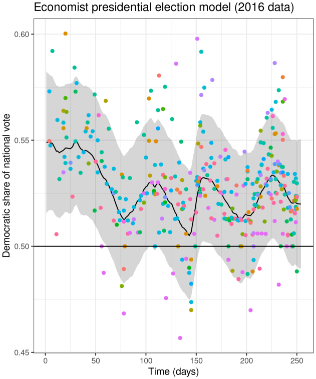


Observations and consequences



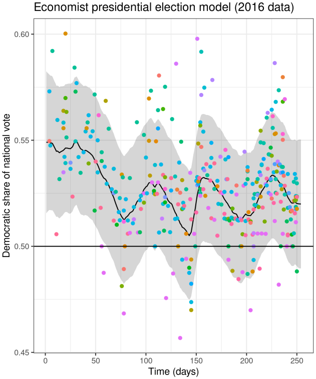
Observations and consequences

- When $\log p(x_n|\gamma, \lambda)$ is the exchangeable unit, our results are problematic for
  - Linear approximations (IJ, AMIP, approx. CV)
  - The nonparametric bootstrap
  - All of the above for Bayes-like optimization procedures (VB, the EM algorithm)



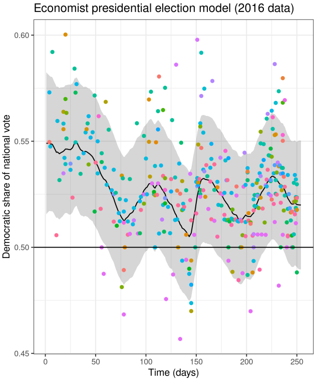
Observations and consequences

- When $\log p(x_n|\gamma, \lambda)$ is the exchangeable unit, our results are problematic for
  - Linear approximations (IJ, AMIP, approx. CV)
  - The nonparametric bootstrap
  - All of the above for Bayes-like optimization procedures (VB, the EM algorithm)
- Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice to be useful.



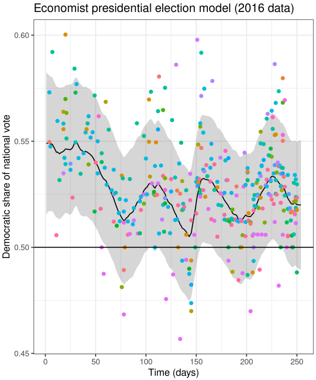
Observations and consequences

- When $\log p(x_n|\gamma, \lambda)$ is the exchangeable unit, our results are problematic for
  - Linear approximations (IJ, AMIP, approx. CV)
  - The nonparametric bootstrap
  - All of the above for Bayes-like optimization procedures (VB, the EM algorithm)
- Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice to be useful.
- There may be multiple ways to define “exchangable unit” in a given problem.



Observations and consequences

- When $\log p(x_n|\gamma, \lambda)$ is the exchangeable unit, our results are problematic for
    - Linear approximations (IJ, AMIP, approx. CV)
    - The nonparametric bootstrap
    - All of the above for Bayes-like optimization procedures (VB, the EM algorithm)
- Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice to be useful.
- There may be multiple ways to define "exchangable unit" in a given problem.
- But without nesting, $\log p(x_n|\gamma, \lambda)$ may be the natural model-free exchangeable unit.



T. Broderick, R. Giordano, and R. Meager. An automatic finite-sample robustness metric: When can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*, 2020.

A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL `https://projects.economist.com/us-2020-forecast/president`. Data and model accessed Oct., 2020.

J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990.

A. Vehtari and J. Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

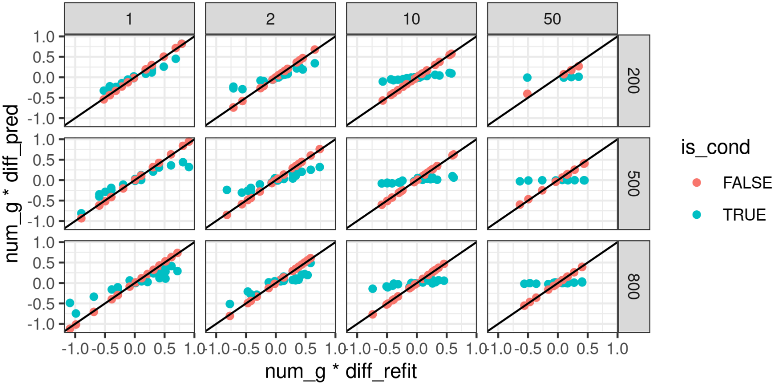
Supplemental slides

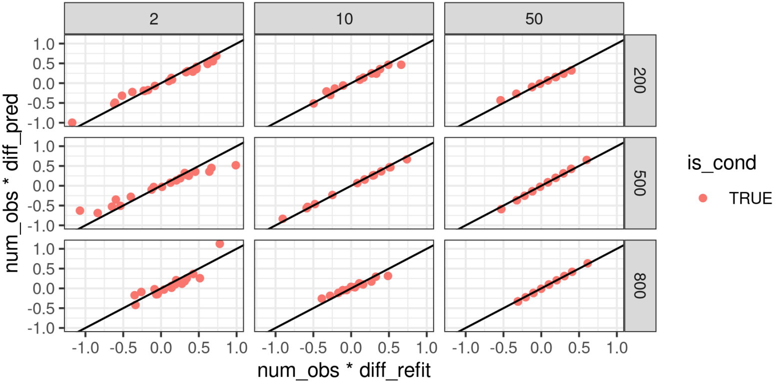
Non-equivalence of weighting (nonlinearity of marginalization)

Consider a single datapoint.

$$
\begin{aligned}
&\log p(x_n|\gamma, w_c) = \\
&\qquad \log\left(\int p(x_n|\gamma,\lambda,w_c)p(\lambda|\gamma)d\lambda\right) = \\
&\qquad \log\left(\int p(x_n|\gamma,\lambda)^{w_c}p(\lambda|\gamma)d\lambda\right) \neq \\
&\qquad \log\left(\int p(x_n|\gamma,\lambda)p(\lambda|\gamma)d\lambda\right)^{w_c} = \\
&\quad w_c \log p(\lambda|\gamma)
\end{aligned}
$$

Extended experimental results




2
10
50
200
500
800
1.0
0.5
0.0
-0.5
-1.0
num_obs * diff_pred
num_obs * diff_refit
is_cond
TRUE